

Augmented images

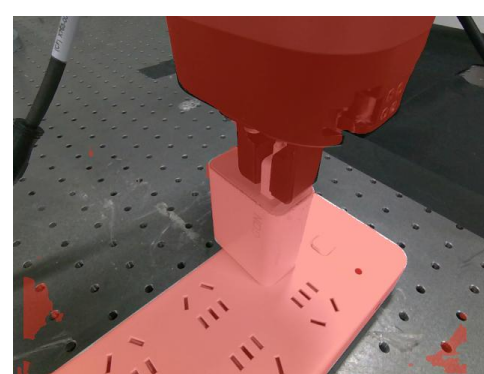


raw image

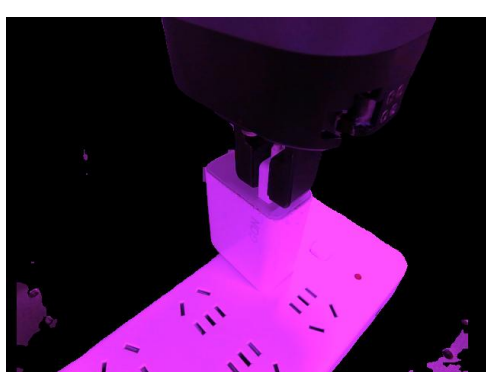
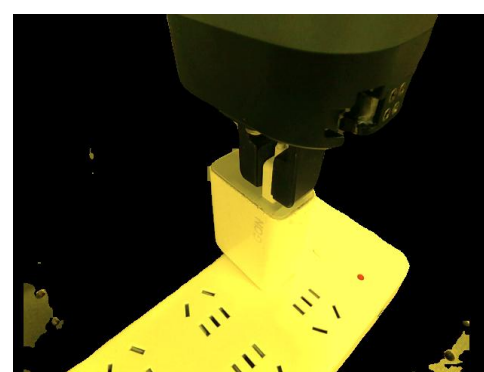
"White charger;
White Strip;
Black Robot"

prompts

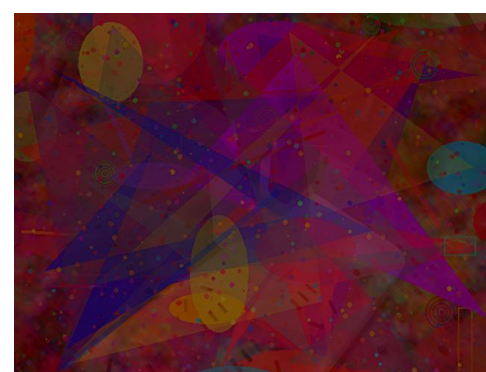
Segmentation Model



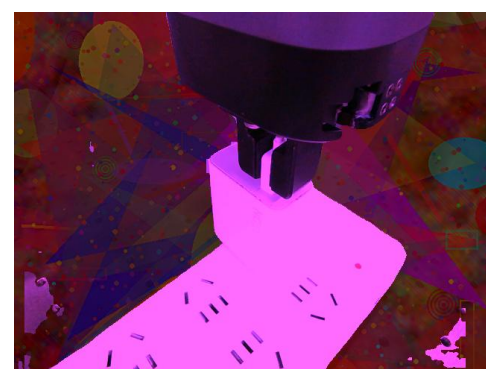
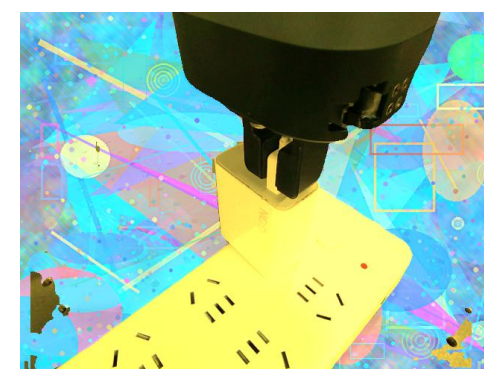
semantic segmentation



(a) random color



(b) random texture



(c) combination

VLM Semantic Localization

VLM Localization



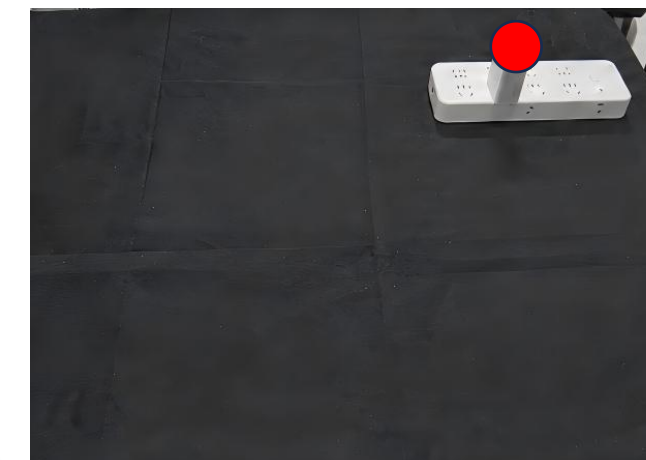
"The USB hole
on the white
charger"

Vision-Language Model

frozen params

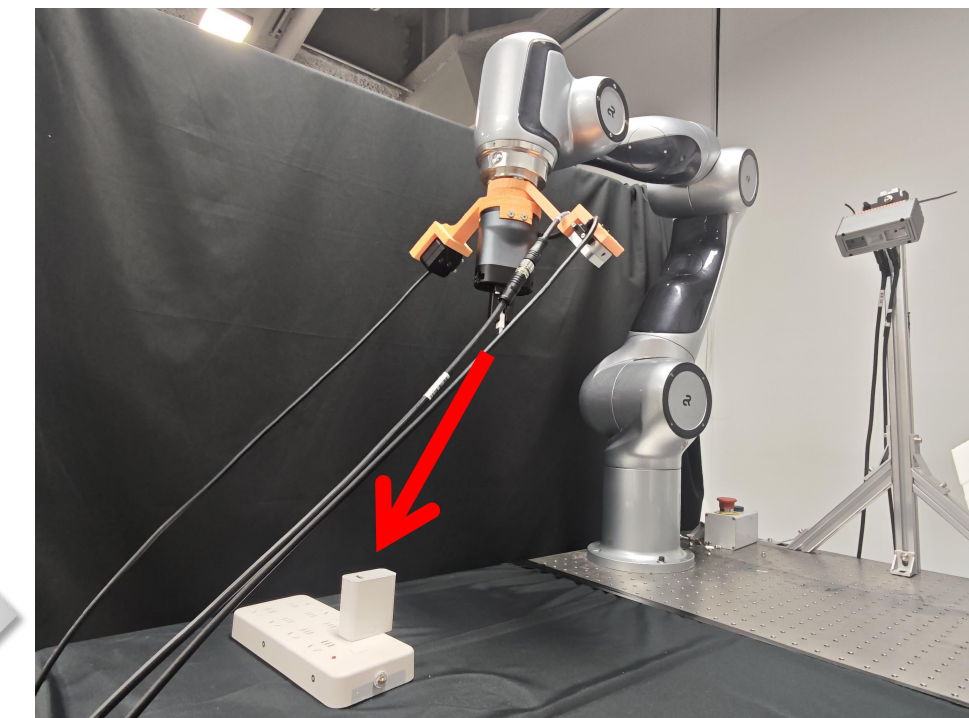
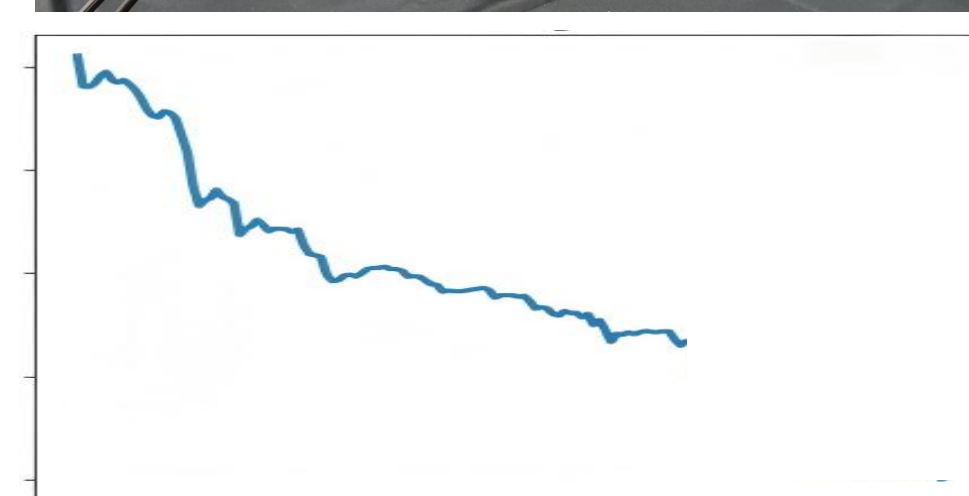
<2D point>
(535.332,
146.845)

3D Calibration



3D
goal

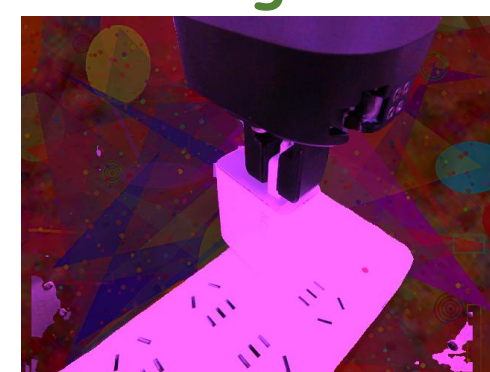
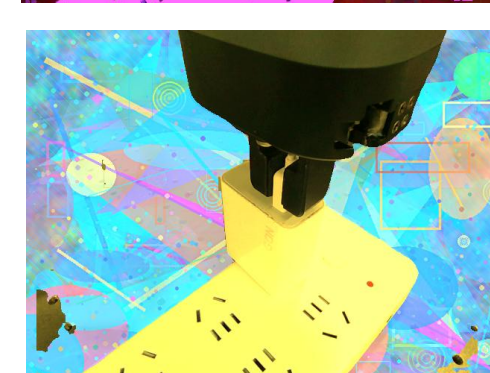
Task progress



VLM planning first completes **80% contact-free** task in progress

RWRL Contact Refinement

Augmented images



Residual Module

Base action chunks

π_{base}

π_{res}

a_{t+1}^{base}
 a_t^{base}

$\oplus$
cat.



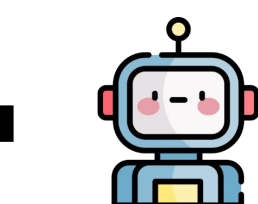
π_{human}

intervene

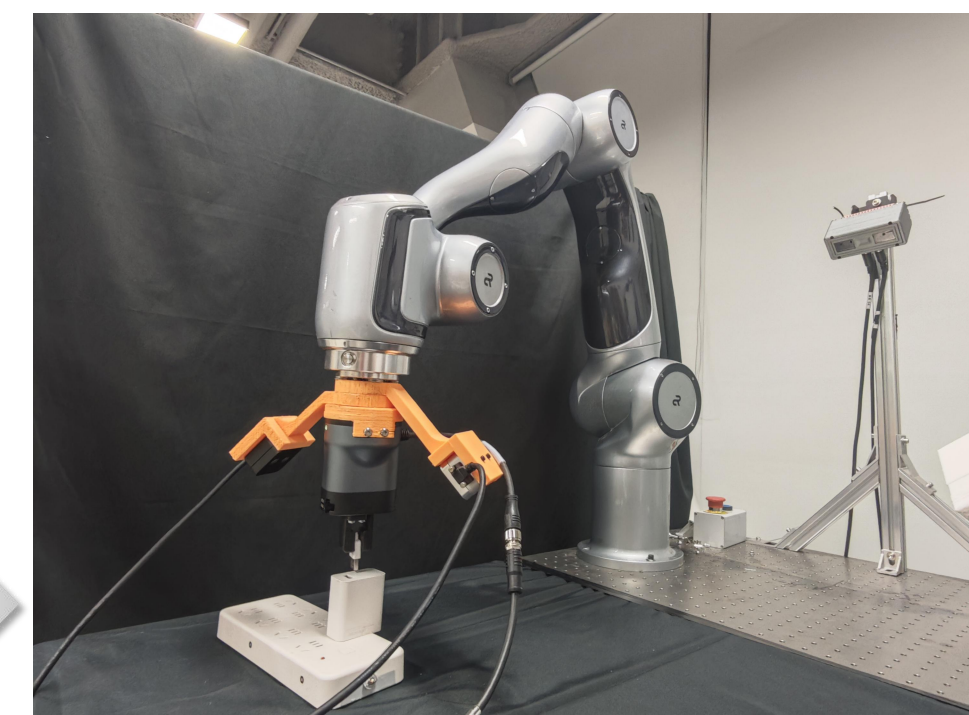
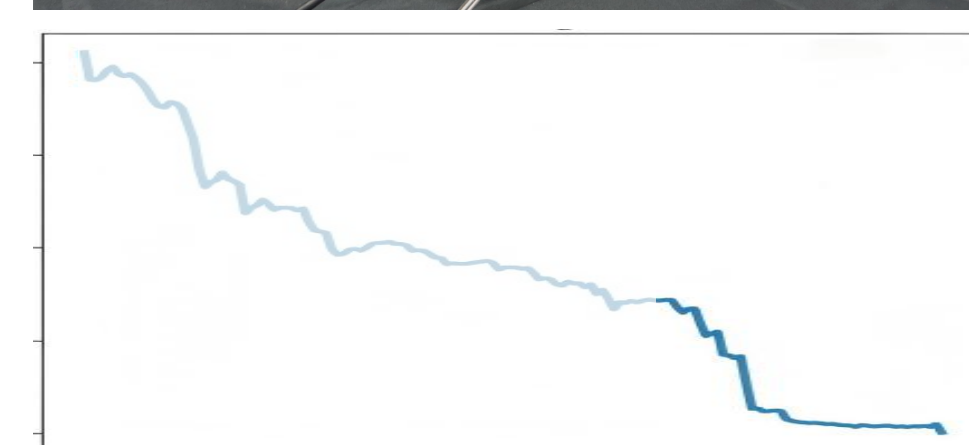
a_{t+1}
 a_t

Offline
buffers
Online
buffers

Real-world environments



Execution



RL agent then completes **20% contact-rich** task in progress